High-resolution genetic maps of *Eucalyptus* improve *Eucalyptus grandis* genome assembly

Jérôme Bartholomé[1,2,3], Eric Mandrou[2,3,4], André Mabiala[5], Jerry Jenkins[6], Ibouniyamine Nabihoudine[4], Christophe Klopp[4], Jeremy Schmutz[6,7], Christophe Plomion[2,3] and Jean-Marc Gion[1,2,3]

[1]CIRAD, UMR AGAP, F-33612 Cestas, France; [2]INRA, UMR1202 BIOGECO, F-33610 Cestas, France; [3]BIOGECO, UMR 1202, Univ. Bordeaux, F-33600 Pessac, France; [4]Plate-forme Bio-informatique Genotoul, INRA, Biométrie et Intelligence Artificielle, BP 52627, 31326 Castanet-Tolosan Cedex, France; [5]CRDPI, BP 1291, Pointe Noire, Republic of Congo; [6]HudsonAlpha Institute for Biotechnology, 601 Genome Way, Huntsville, AL 35801, USA; [7]US Department of Energy Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA 94598, USA

Summary

- Genetic maps are key tools in genetic research as they constitute the framework for many applications, such as quantitative trait locus analysis, and support the assembly of genome sequences.
- The resequencing of the two parents of a cross between *Eucalyptus urophylla* and *Eucalyptus grandis* was used to design a single nucleotide polymorphism (SNP) array of 6000 markers evenly distributed along the *E. grandis* genome.
- The genotyping of 1025 offspring enabled the construction of two high-resolution genetic maps containing 1832 and 1773 markers with an average marker interval of 0.45 and 0.5 cM for *E. grandis* and *E. urophylla*, respectively. The comparison between genetic maps and the reference genome highlighted 85% of collinear regions. A total of 43 noncollinear regions and 13 nonsynthetic regions were detected and corrected in the new genome assembly. This improved version contains 4943 scaffolds totalling 691.3 Mb of which 88.6% were captured by the 11 chromosomes. The mapping data were also used to investigate the effect of popula-tion size and number of markers on linkage mapping accuracy.
- This study provides the most reliable linkage maps for *Eucalyptus* and version 2.0 of the *E. grandis* genome.

A century after the publication of the first genetic map of *Drosophila melanogaster* (Sturtevant, 1913), linkage maps have been established for many organisms (Botstein *et al.*, 1980; Koornneef *et al.*, 1983; Grattapaglia & Sederoff, 1994) and are central to genetic research. Based on the recombination events between homologous chromosomes occurring during meiosis (i.e. crossover in prophase I; Crismani *et al.*, 2013), linkage maps provide a linear representation of the order and the distance between markers on chromosomes. Since the arrival of molecular markers (Schlotterer, 2004) and recent technological and bioinformatic advances facilitating their high-throughput discovery (Kumar *et al.*, 2012), genetic maps have enlarged their applications. Indeed, genetic maps constitute a powerful tool for characterizing the genome structure of nonsequenced species and play a major role in genome assembly and validation (e.g. scaffold anchoring) as illustrated in *Arabidopsis thaliana* (Kaul *et al.*, 2000), *Homo sapiens* (Lander *et al.*, 2001) and *Populus trichocarpa* (Tuskan *et al.*, 2006). Genetic maps also provide insights into genome evolution through the analysis of synteny, collinearity and chromosomal rearrangements between species (Burt, 2002; Choi *et al.*, 2004; Krutovsky *et al.*, 2004; Hudson *et al.*, 2012b). Moreover, linkage maps form the basis for mapping and cloning quantitative trait loci (QTL) involved in the genetic control of traits of interest (Price, 2006; Salvi & Tuberosa, 2007; Mackay *et al.*, 2009; Wurschum, 2012).

The precision of estimates of recombination rates between linked markers, which ultimately determines map accuracy, is crucial for all of these applications. Several mapping algorithms have been developed to order the ever-increasing number of markers and estimate genetic distances with improved speed and reliability (Cheema & Dicks, 2009; Mollinari *et al.*, 2009). Merging methods have also been proposed to combine linkage maps (Stam, 1993; Peirce *et al.*, 2007; Wu *et al.*, 2011; Endelman & Plomion, 2014) and build composite genetic maps. Although mapping algorithms have been shown to play an important role in mapping accuracy (Collard *et al.*, 2009; Mollinari *et al.*, 2009; Ronin *et al.*, 2010), it is primarily the number of recombination events captured in the mapping population, depending on the population type and its sample size, that determines mapping accuracy (Nelson, 2005; Ferreira *et al.*, 2006). In order to build

high-resolution genetic linkage maps with a practical sample size, highly recombinant populations, for example intermated recombinant inbred lines (Ganal *et al.*, 2011), and sperm or pollen typing techniques (Yelina *et al.*, 2012) were developed. In outbred species, such as forest trees, mapping activities have mostly relied on existing full-sib families of rather small sample sizes (Kole, 2007), thereby limiting the accuracy and resolution of genetic maps. Such families were generated as components of breeding programmes for genetic parameter estimation. Although the number of markers is no longer a limiting factor (Sansaloni *et al.*, 2010; Geraldes *et al.*, 2013; Howe *et al.*, 2013) for building high-density linkage maps in forest trees, there is still room for progress in improving mapping accuracy with a larger population size, although at the expense of genotyping costs.

Genetic mapping approaches have been widely used to characterize the unsequenced genomes of many forest tree species over recent decades (Kole, 2007; Neale & Kremer, 2011). For *Eucalyptus*, the most widely grown plantation hardwoods, numerous genetic maps have been built for different species (reviewed in Supporting Information Table S1). The first genetic maps were developed using an interspecific cross between *Eucalyptus grandis* × *Eucalyptus urophylla* with dominant markers (Grattapaglia & Sederoff, 1994). Then, the development of simple sequence repeat markers (Brondani *et al.*, 1998, 2006), as well as the identification of expressed sequenced tag (EST) polymorphisms (Gion *et al.*, 2000; Thamarus *et al.*, 2002), enabled broad-scale comparisons of genome organization across species (J.-M. Gion *et al.*, unpublished). More recently, Diversity Arrays Technology (DArT) markers were developed for the *Eucalyptus* and *Corymbia* species (Sansaloni *et al.*, 2010). Genetic maps with up to 4000 DArTs for *E. urophylla*, *E. grandis* and *Eucalyptus globulus* (Hudson *et al.*, 2012a,b; Kullan *et al.*, 2012; Petroli *et al.*, 2012) were constructed, specifying linkage group homologies as well as macro-synteny and collinearity between these species. However, most of the linkage maps were established with a small to moderate number of full-sib or backcross offspring, ranging from 62 (Grattapaglia & Sederoff, 1994) to 503 (Hudson *et al.*, 2012b). Recently, a new era began for eucalyptus genomics with two main breakthroughs: (1) physical characterization of the *E. grandis* BRASUZ1 reference genome (Myburg *et al.*, 2014) totalling 691 Mb distributed over 4952 scaffolds; the main 11 super-scaffolds, representing the 11 chromosomes (Chr), accounted for 88% (605.9 Mb) of the genome and for 93% (33 917) of the predicted genes; and (2) the development of a vast number of polymorphisms for different eucalyptus species (Novaes *et al.*, 2008; Sansaloni *et al.*, 2010; Grattapaglia *et al.*, 2011; Neves *et al.*, 2011), providing a considerable number of markers for high-density linkage mapping in this genus.

In this study, we developed genomic resources for the two parents of an interspecific cross between *E. urophylla* and *E. grandis* by resequencing their whole genomes. The large number of single nucleotide polymorphisms (SNPs) detected after the mapping of paired-end reads on the BRASUZ1 genome enabled us to design a high-quality SNP genotyping array maximizing the coverage of the 11 chromosomes. Genotyping of these markers on a large full-sib family totalling 1025 F1s made it possible to achieve our three initial objectives: to construct two high-resolution linkage maps; to test the effect of population size and number of markers on genetic mapping accuracy; and to analyse the synteny and collinearity between the genetic maps and the BRASUZ1 sequence in order to improve the first version of the *E. grandis* genome.

Materials and Methods

Plant material and DNA extraction

We used an interspecific cross between *E. urophylla* S.T. Blake as a female parent and *E. grandis* W. Hill ex Maiden as a male parent. These two species are phylogenetically related as they belong to the same subgenus (*Symphyomyrtus*) and the same section (*Latoangulatae*) (Steane *et al.*, 2002, 2011). The F1 mapping population included 1025 full-sibs planted in a field trial using seedlings. For the genome sequencing of the two parents, total genomic DNA was extracted from frozen leaves using a modified protocol from Doyle & Doyle (1990). For SNP genotyping of the two parents and the 1025 offspring, DNA was extracted from dry leaves with a protocol using magnetic beads (Smart D-N-Adem-Kit from Ademtech, Pessac, France). To assess DNA quality and quantity, the following analyses were carried out on each genotype: electrophoresis on 0.8% agarose gel and quantification by both spectrophotometry (NanoDrop 8000; Thermo Scientific, Wilmington, DE, USA) and fluorescence (Quant-iT Picogreen; Invitrogen, Carlsbad, CA, USA).

SNP detection and array design

Whole-genome resequencing was performed with an Illumina HiSeq2000 (GATC Biotech AG, Konstanz, Germany) for the two parental genotypes (*E. urophylla* and *E. grandis*). Two libraries were constructed for each parent: a 300-bp fragment size library (100 bp paired-end) and a 3-kb mate pair fragment size library (50 bp paired-end). All steps from sequencing to array design are summarized in Fig. 1.

The *E. grandis* BRASUZ1 genome assembly v1.0 (Myburg *et al.*, 2014) was used as the reference genome for mapping the short reads. Mapping was achieved independently for the two parental genotypes using the Burrows–Wheeler Alignment tool (BWA, version 0.6.1-r104 (Li & Durbin, 2009)) with standard parameters. Then, the Genome Analysis Toolkit (GATK, version 1.5-21-g979a84a (McKenna *et al.*, 2010)) was used to realign insertion/deletion (INDEL) sites and recalibrate PHRED quality scores (CountCovariates, IndelRealigner and TableRecalibration procedures with standard parameters). At the end of this step, one set of aligned short reads was available for each parent, combining 50- and 100-bp reads. Only sequences with high mapping quality (MAPQ score ≥ 30) were used in the next steps.

Variant detection was performed using GATK (UnifiedGenotyper procedure with options – standcallconf 50.0 and – standemitconf 10.0) pulling together the two parental sets of aligned sequences. Different filters were applied to the initial set of SNPs in order to obtain high-quality *in silico* SNPs and design a 6K

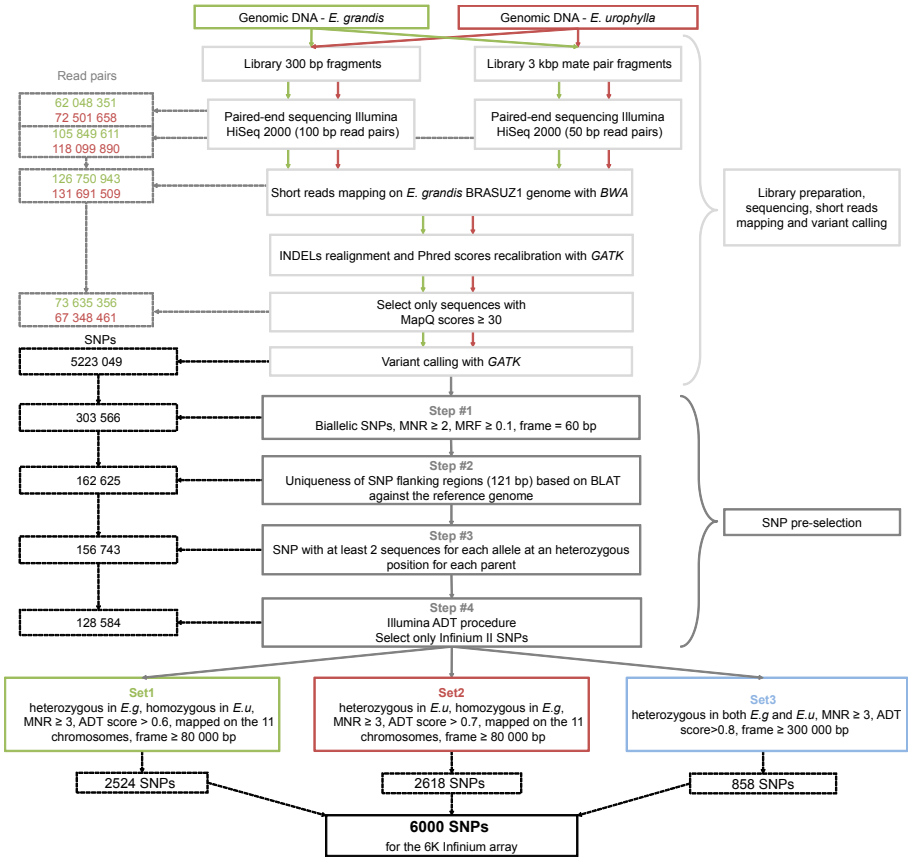

Fig. 1 Flowchart for Infinium single nucleotide polymorphism (SNP) array design. ADT, Assay Design Tool; INDEL, insertion/deletion; MAPQ, mapping quality; MFR, minimum read frequency; MNR, minimum number of reads.

array (Fig. 1). SNP pre-selection was divided into four steps. First, only bi-allelic SNPs (within or between parents) were kept and INDEL variants were discarded. Selection of these SNPs was based on a minimum number of reads (MNR) per allele (MNR ≥ 2), a minimum read frequency (MRF) between the two alleles (MRF ≥ 0.1) and a minimum distance of 60 bp between adjacent SNPs (Fig. 1, Step #1). Secondly, the 121-bp sequence surrounding the SNP (i.e. the SNP and its two 60-bp flanking regions) was extracted and checked for its uniqueness against the BRASUZ1 genome using BLAT software (Kent, 2002). SNP flanking sequences associated with more than one hit against the genome were discarded (Step #2). Then (Step #3), for a given SNP, each parental allele had to be represented by at least two reads. Finally (Step #4), a designability score, based on the Illumina Assay Design Tool (ADT), was calculated (Shen *et al.*, 2005). To maximize the number of SNPs on the array, only Infinium type II SNPs were used.

Three categories of SNPs were finally selected based on their expected segregation patterns in the full-sib family (Set 1, Set 2 and Set 3 as defined in Fig. 1). Set 1 consisted of SNPs heterozygous in *E. grandis* and homozygous in *E. urophylla* (segregating in a 1 : 1 Mendelian ratio), located on the 11 chromosomes with a minimum distance between two consecutive SNPs of 80 kb, an MNR ≥ 3 and an ADT score > 0.6. Set 2 comprised SNPs heterozygous in *E. urophylla* and homozygous in *E. grandis* (segregating 1 : 1). As a larger number of informative markers was initially available for Set 2, the same criteria as for Set 1 were used with a more stringent ADT score (> 0.7). In Set 3, SNPs were heterozygous in both parents (segregating 1 : 2 : 1). No restriction on scaffold size was applied for Set 3. For the other criteria a minimum interval between SNPs of 300 kb, an MNR ≥ 3 and an ADT score > 0.8 were used. These three sets were used to design a 6K Infinium array (Illumina, San Diego, CA, USA).

SNP genotyping

SNPs were genotyped by PEGASE-biosciences (Douai, France). The results were analysed with GENOMESTUDIO (Genotyping module V1.9; Illumina). Poorly performing individuals were removed from the analysis when they were lower than the following thresholds: 0.49 for the 10% GENCALL score or 0.98 for the call rate. The clustering of each SNP was visually checked for its relevance to the expected inheritance pattern based on parental genotypes (Fig. S1). When a cluster was not reliable, based on 24 replicates of each parent, it was re-clustered manually. SNPs with a low fluorescence intensity, a call frequency $> 0.90\%$ or a GENTRAIN score > 0.4 were discarded. The reproducibility of the genotyping was evaluated using 24 replicates for each parent and two replicates for three offspring. All successful SNPs are shown in Table S2.

Genetic linkage analysis

Construction of genetic linkage maps The linkage analysis was performed with R v3.0.1 using ONEMAP v2.0-3 (Margarido *et al.*, 2007; Mollinari *et al.*, 2009). Genetic maps were constructed for

each parental genotype (*E. grandis* and *E. urophylla*) according to a two-way pseudo-test-cross mapping strategy (Grattapaglia & Sederoff, 1994). The two parental maps were built independently with the same procedure. For linkage analysis we used SNPs and individuals with < 4% and < 2% missing data, respectively. For the two parental maps, SNPs were grouped into linkage groups (LGs) with a stringent threshold (logarithm of odds (LOD) score ≥ 30). In the first step, marker ordering within LGs was performed with the RECORD algorithm (Os *et al.*, 2005) implemented in ONEMAP using only SNPs segregating 1 : 1 (Set 1 and Set 2, test-cross markers) to build the framework maps (LOD score threshold = 3). To validate the marker order obtained with RECORD, framework maps were compared to those obtained with the maximum likelihood (ML) algorithm implemented in JOINMAP V4.1 (Van Ooijen, 2011b). In the second step, SNPs segregating 1 : 2 : 1 (Set 3, inter-cross markers) were added to the framework maps using the RECORD algorithm. Genetic distances (cM) were calculated using the Haldane mapping function (Haldane, 1919) in order to compare the RECORD and the ML algorithm, as the ML algorithm of JOINMAP only uses the Haldane mapping function. SNP segregations were tested for goodness of fit to the expected Mendelian segregation ratios using χ^2 tests with the level of significance adjusted for simultaneous multiple tests (Benjamini & Yekutieli, 2001) within each LG of the parental maps.

Linkage mapping accuracy The effect of the sample size (number of offspring) and marker density on linkage map accuracy was tested based on re-sampling in the whole data set. Three sample sizes (100, 200 and 500 individuals) and four marker densities (using only markers segregating 1 : 1) were used. For each sample size, a random draw (1000 times) was performed from the 1020 offspring. Marker selection was based on genetic distances obtained with the framework maps. Four density classes were selected: one marker every 5, 2.5 and 1.2 cM and all markers (referred to as MD1, MD2, MD3 and MD4, respectively). In total, 12 000 genetic maps for each parent (3 sample sizes × 4 marker densities × 1000 samples) were built using the RECORD algorithm of ONEMAP. For each map, grouping into LGs was based on the framework maps.

Genetic map integration and genome assembly

The two parental framework maps were used to detect putative false joins within and between scaffolds in the original assembly (Myburg *et al.*, 2014). Scaffolds were broken if they contained a false join (indicated by more than one marker) coincident with an area of low bacterial artificial chromosome/fosmid coverage. Internal utilities in ARACHNE (Jaffe *et al.*, 2003) were used to perform scaffold breaks, preserving the underlying alignment and assembly information. Markers and annotated exons from *E. grandis* genome v1.0 were aligned to the broken scaffolds to make sure no information was lost during the process. Telomeric sequences were identified using the $(TTTAGGG)_n$ repeat, and care was taken to make sure that it was properly oriented in the version 2.0 assembly. Optimal order and orientation of the scaffolds were obtained using the genetic maps. Each map join between two scaffolds was sized with 10 000 Ns. In the absence of marker evidence, scaffolds present in the first version of the genome were retained in their original order and orientation.

Scaffolds were classified into bins depending on sequence content. Contamination was identified using MEGABLAST against the NCBI nucleotide collection (NR/NT) and BLASTX against a set of known microbial proteins (Zhang *et al.*, 2000). The completeness of the euchromatic portion of the release assembly was assessed by aligning 1.6 million 454 EST sequences to the genome.

Results

Resequencing, SNP selection and SNP array design

Considering all libraries, resequencing of the two parental genotypes resulted in 49.3 Gb of sequences. A total of 62 048 351 (72 501 658) 100-bp read pairs and 105 849 611 (118 099 890) 50-bp read pairs were obtained for *E. grandis* (*E. urophylla*). According to the estimated physical size of the *E. grandis* (640 Mb) and *E. urophylla* (650 Mb) genomes (Grattapaglia & Bradshaw, 1994), the theoretical mean haploid genome coverage was 35.9X and 40.5X for *E. grandis* and *E. urophylla*, respectively. The proportion of aligned reads on the BRASUZ1 genome was higher for *E. grandis* (75.5%) than for *E. urophylla* (69.1%), as might be expected given the species of the reference genome (*E. grandis*) and the small phylogenetic distance between *E. grandis* and *E. urophylla*. The distribution of aligned reads along the genome was homogenous between chromosomes and between read types: 100 or 50 bp (Fig. S2). Only robustly aligned reads (MapQ > 30) were used for SNP detection, that is, 51.1% (*E. grandis*) and 58.1% (*E. urophylla*) of the aligned reads.

The variant calling procedure with GATK resulted in 5 223 049 SNPs being polymorphic either within or between the two parental genotypes. After applying stringent selection criteria to the initial set of detected SNPs (Fig. 1, Step #1 to #4), 128 584 high-quality *in silico* SNPs were obtained. Lastly, we selected three complementary sets of SNPs based on their comprehensive distribution along the genome and on their expected segregation patterns in the progeny (Fig. S3), as follows: Set 1 presented 2524 SNPs heterozygous in *E. grandis* only (1 : 1), Set 2 presented 2618 SNPs heterozygous in *E. urophylla* only (1 : 1 segregation) and Set 3 presented 858 SNPs heterozygous in both parents (1 : 2 : 1). The selected SNPs were evenly spaced along the BRASUZ1 genome with an average physical distance between adjacent SNPs of 100 kb (± 163 kb). However, some regions amounting to 8% of the chromosome length on average were less well covered as a consequence of a distance of > 1 Mb between adjacent markers. Most of the time these regions were characterized by a high rate of missing data (gaps) in the genome (e.g. 25% from 20 to 22 Mb on Chr3 and 23% from 32 to 34 Mb on Chr5; Fig. S3).

Genotyping quality

The success rate for custom BeadChip manufacturing was 86.25%, that is, 5175 SNPs were available for genotyping. After SNP clustering and quality control using GenomeStudio, five genotypes with a low call rate (< 0.98) or a low 10% GenCall (< 0.49) were removed from further analysis. In all, 339 SNPs (6.5%) with technical failures (i.e. no call) were discarded (Table 1). Of the remaining 4836 SNPs, 469 were monomorphic and 4367 exhibited clear segregation within the full-sib family, giving a genotyping success rate (SR) of 84.4%. Although the average sequencing depth for all genotyped SNPs was similar between *E. grandis* and *E. urophylla*, the proportion of monomorphic markers was lower for the former (3.9%) compared with the latter (15.9%). This must have been related to the differences in genome structure between the two species and the less optimum alignment of *E. urophylla* short reads on the *E. grandis* genome. Moreover, the distribution of monomorphic markers between chromosomes was different between the two parental sets (Table S3). Within chromosomes, some regions from 1 to 3.9 Mb were found to gather monomorphic SNPs for one parent and polymorphic SNPs for the other: on Chr6 for *E. grandis* and on Chr4, 5, 6 and 10 for *E. urophylla* (Table S4). For the polymorphic SNPs, the average call frequency was 99.8% (± 0.58%), the average GenTrain score was 79.9% (± 6.3%) and the average cluster separation score was 0.91 (± 0.18). Considering the three SNP sets, the average physical distance between adjacent polymorphic SNPs was 138 kb (Fig. 2a). Taking the sets separately, the average marker interval was larger: 321 kb for Set 1, 332 kb for Set 2 and 797 kb for Set 3 (Fig. 2b). Genotyping reproducibility over all polymorphic SNPs was at least 99.95%. Using version 1.1 of the predicted gene models of the *E. grandis* genome, 38% of the polymorphic SNPs were found to map into genes (Set 1: 32%; Set 2: 39.5%, and Set 3: 52.1%). Moreover, 78% of the polymorphic SNPs were located within 5 kbp of predicted gene models.

Genetic mapping

Construction of high-resolution genetic linkage maps The maps were established using the same procedure and independently for the two parents. Set 1 and Set 2 (SNPs segregating 1 : 1) were first used for the construction of two framework maps. The total map length was lower for *E. grandis* (821.7 cM) than for *E. urophylla* (885.9 cM) (Tables 2, S5). Framework maps provided a high resolution with one marker every 0.45 cM (*E. grandis*) and 0.50 cM (*E. urophylla*) on average. Despite greater coverage of the reference genome for *E. grandis* maps (98.2%) than for *E. urophylla* maps (97.2%), the *E. grandis* map was significantly shorter compared with that of *E. urophylla*. Moreover, this difference was underestimated because of the low coverage of Chr4 for *E. urophylla* (85.6%) resulting in a shorter LG4 length (5.5 cM) compared with *E. grandis*. Overall, no particular trend in terms of LG length was found between species (Table S5). SNPs with significant segregation distortion ($P_{adjust} < 0.01$) were kept for linkage map construction and accounted for 21% and 30.7% of the mapped SNPs for *E. grandis* and *E. urophylla*, respectively (Table 2). SNPs with segregation distortion occurred in large regions, the so-called segregation distortion region (SDRs) on several LGs with differences between parental maps (Fig. S4). LG1 and 3 presented a high level of distortion in *E. grandis* only, LG5, 6, 7 and 11 did so only in *E. urophylla*, LG2 presented a high level of distortion in both parental maps and LG4, 8, 9 and 10 presented no distortion in either parental map (Fig. S4).

In a second step, 719 SNPs from Set 3 (segregating 1 : 2 : 1) were added to the framework maps. The resulting maps comprised 2552 and 2493 SNPs for a total map length of 912.6 cM for *E. grandis* and 904 cM for *E. urophylla* (Tables 2,S6), corresponding to an additional 90.9 and 18.1 cM compared with the framework maps, for *E. grandis* and *E. urophylla*, respectively. The resolution provided by these two maps was high, with one marker every 0.36 cM. Conversely to the comparison between parental framework maps, the total map length was slightly greater for *E. grandis* (8.3 cM longer) than for *E. urophylla*, probably as a result of less optimum coverage of Chr4 for *E. urophylla*. Thus, the integration of SNPs from Set 3 tended to hide the difference in map length between the two parental framework maps. Moreover, the density of inter-cross markers every 1.26 cM (*E. grandis*) and 1.34 cM (*E. urophylla*) enabled clear identification of orthologous regions between *E. grandis* and *E. urophylla* parents. The order of inter-cross markers between parental maps was well conserved, except for 9.3% of them, mainly located on LG2, 3, 5 and 7 (Fig. S5). These inversions (0.4 cM on average) were probably related to a loss of accuracy in the estimation of the recombination rate between pairs of markers segregating 1 : 1 and 1 : 2 : 1 (Ritter *et al.*, 1990). Indeed, only half of the genotypes were informative for the estimation of the recombination rate between test-cross and inter-cross markers.

Effect of sample size on mapping accuracy The effect of sample size and marker density (MD) on genetic mapping accuracy was tested using a random draw in the whole data set. MD classes were based on framework maps (Table S7). The results obtained were highly similar for both parental maps. Regardless of MD, the sample size had little impact on the average map length (Fig. 3a; Table S8), with a general trend towards longer map lengths for small sample sizes (e.g. for MD1 the average map length (*E. urophylla*) was 921.8, 917.1 and 913.9 cM for $n = 100$, 200 and 500, respectively). The combined effect of sample size

Table 1 Genotyping results: distribution of polymorphic, monomorphic and failed (i.e. no call) single nucleotide polymorphisms (SNPs) for the three sets presented in the Materials and Methods section: Set 1 and 2 test-cross markers informative for the *Eucalyptus grandis* and *Eucalyptus urophylla* parents, respectively; Set 3 inter-cross markers informative in both parents

	Polymorphic (%)	Monomorphic (%)	No call (%)	Total
Set 1	1849 (88.6)	82 (3.9)	157 (7.5)	2088
Set 2	1793 (78)	366 (15.9)	139 (6)	2298
Set 3	725 (91.9)	21 (2.7)	43 (5.4)	789
Total	4367 (84.4)	469 (9.1)	339 (6.6)	5175

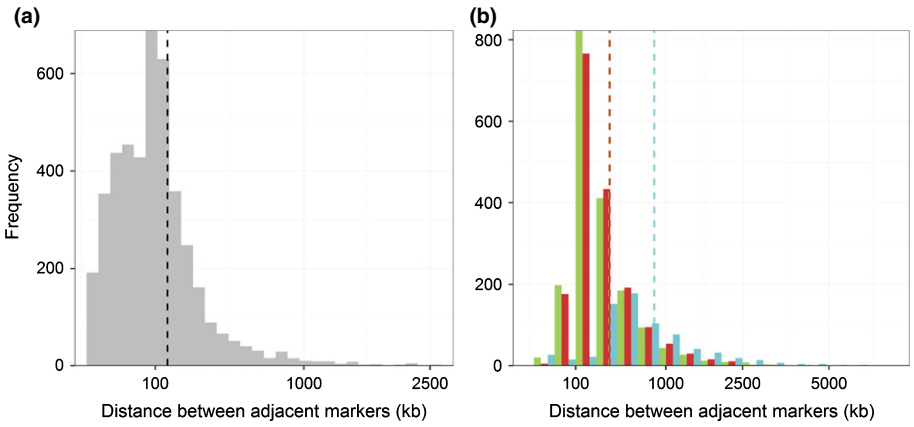

Fig. 2 Distribution of inter-marker distances for the polymorphic single nucleotide polymorphisms (SNPs) for (a) the whole data set and (b) the three SNP sets (see the Materials and Methods section). Set 1, green; Set 2, red; Set 3, blue. The vertical dashed lines represent the mean distance between adjacent markers (321 kb for Set 1, 332 kb for Set 2 and 797 kb for Set 3).

Table 2 Characteristics of framework (test-cross markers) and complete linkage maps (test-cross and inter-cross markers) for *Eucalyptus grandis* (*E.g*) and *Eucalyptus urophylla* (*E.u*)

	Framework map		Complete map	
	E.g	*E.u*	*E.g*	*E.u*
Map length (cM)	821.66	885.92	912.59	903.99
Number of SNPs (total)	1832	1773	2551	2491
Number of SNPs (1 : 1)	1832	1773	1832	1773
Number of SNPs (1 : 2 : 1)	–	–	719	718
Number of genetic bins	1429	1353	1904	1799
Distance between SNPs (cM)	0.454	0.498	0.358	0.362
Distance between genetic bins (cM)	0.578	0.652	0.48	0.5
Distorted SNPs (%)	21	30.7	19.8	28.7
Linkage map coverage (%)	98.2	97.2	98.3	97.3

SNPs, single nucleotide polymorphisms.
Genetic bin, position on a genetic map defined by one or more linked markers.

and MD was more significant on the number of genetic bins (Fig. 3b) and the average distance between bins (Fig. 3c). Indeed, for low MD the sample size had little effect on the number of bins, and hence average distances between genetic bins (*E. urophylla*, MD1) were 5, 4.9 and 4.9 cM for $n = 100$, 200 and 500, respectively. In contrast, for a higher MD, the increase in sample size improved the resolution of the maps. For MD4, the number of bins was doubled with 500 samples compared with 100 samples (1135.5 bins versus 560.6; Table S8). However, the increase in MD was also related to a decrease in accuracy, as highlighted by the larger number of inverted bins and the decrease in Spearman's rank correlation coefficient (Fig. 3d). For all map features, greater variability was observed for a small sample size (100 or 200) compared with the sample size of 500, reflecting the high variability between samples of the same family.

Comparison between genetic maps and the *E. grandis* genome

The high resolution in marker ordering made it possible to compare SNP locations on the *E. grandis* and *E. urophylla* framework maps with their respective position on the genome (Fig. 4). Overall, the two high-resolution linkage maps displayed a high level of collinearity with the reference genome, that is, conserved the order of SNPs between their genetic and physical locations. These collinear regions amounted to 85.7% of the 605.9 Mb of the 11 chromosomes for *E. grandis* and 84.1% for *E. urophylla*.

Discordant genetic regions were also highlighted (a discordant region is defined by at least two successive mapped SNPs) in comparison to the genome assembly. We distinguished: differences in region ordering and/or orientation within chromosomes, called noncollinear regions (NCRs); and nonconserved chromosomal assignments between genetic and physical maps, called nonsyntenic regions (NSRs). In all, 49 and 51 NCRs were identified for *E. grandis* and *E. urophylla*, representing 13.9% and 15.3% of the genome size (the 11 chromosomes), respectively (Table 3; Fig. 5). Of the 100 NCRs (Table S9), 43 concerned nine out of the 11 chromosomes, and were consistent between the two parental maps. They represented 13.5% (*E. grandis*) and 14.4% (*E. urophylla*) of the genome. Considering only these common NCRs, Chr1 presented the largest differences between the physical and genetic maps, with NCRs accounting for 46.4% (*E. grandis*) and 43.3% (*E. urophylla*) of the chromosome size (Table S9). The chromosome with the second largest proportion of common NCRs was Chr5, in which they accounted for 22.8% (*E. grandis*) and 27.1% (*E. urophylla*) of its size. Six and eight specific NCRs were also found for *E. grandis and E. urophylla*, respectively. They were mainly related to a smaller number (or the absence) of SNPs in one parent compared with the other. These specific NCRs were distributed over Chr1, 3 and 7 for *E. grandis* and Chr1, 2, 3, 6, 7 and 9 for *E. urophylla*. For both parental maps, Chr7 only presented specific NCRs. To a lesser extent, nonconserved chromosomal assignments (NSRs) were revealed by the comparison between the physical and genetic maps (Fig. 5 and Table 3). With the exception of Chr3, we identified 27 NSRs on all the other chromosomes which represented 1% (*E. grandis*) and 0.7% (*E. urophylla*) of the targeted chromosome size, of which 13 were common to the two parental maps (Table S10). Considering common NSRs, Chr2 presented the largest NSR, representing 2.9% (*E. grandis*) and 3.5% (*E. urophylla*) of the chromosome size. Specific NSRs were also found on Chr3, 6, 7

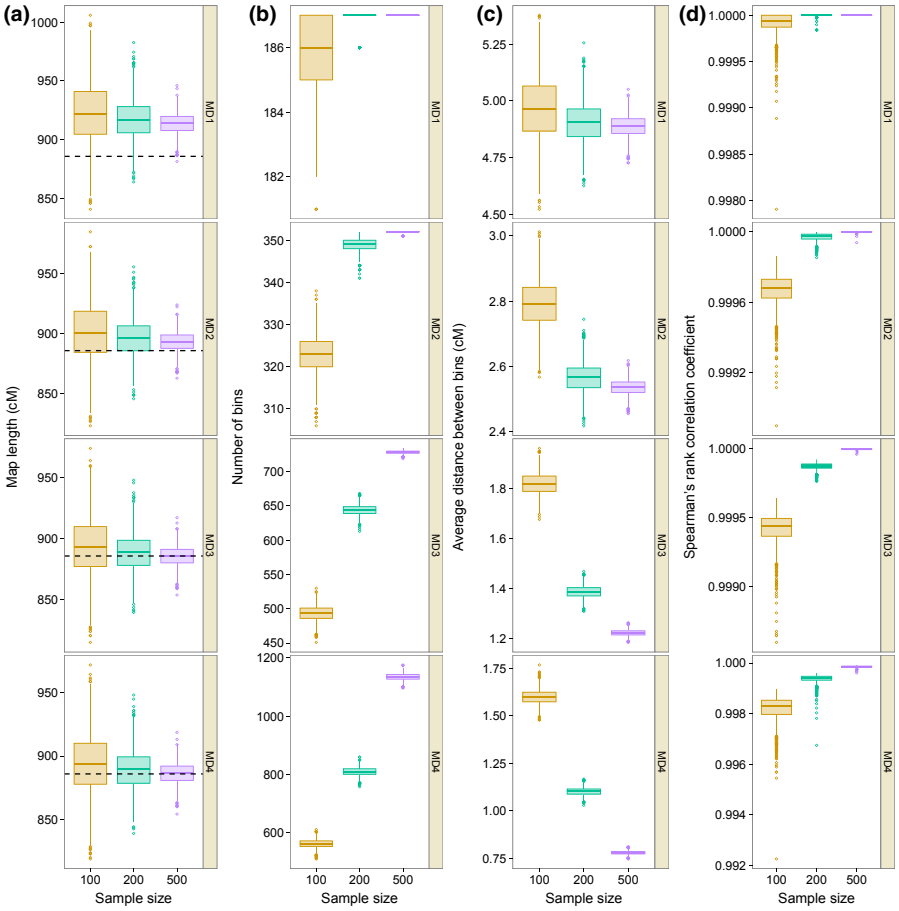

Fig. 3 Boxplot of map length (a), numbers of genetic bins (b), average distance between genetic bins (c) and Spearman's rank correlation coefficient (d) for different combinations of marker density (MD) and sample size resulting from 1000 maps for each combination. The horizontal dashed lines represent the length of the framework maps for *Eucalyptus urophylla* (*E.u*).

and 9. As explained previously for NCRs, specific NSRs were mainly related to the absence of SNPs in the other parental map, but in four cases (on Chr6, 7 and 8 for *E. grandis* and on Chr9 for *E. urophylla*) the closest marker in the orthologous map was well localized (Table S10). This could be explained by the presence of paralogous SNPs.

Improvement of the *E. grandis* genome assembly

The NCRs and NSRs that were corroborated by the two parental genetic maps strongly suggested false joins in the first version of the *E. grandis* genome assembly. Therefore, genetic maps jointly with sequence information were used to identify and break a total of 41 false joins mainly located on Chr1 and Chr6, with nine and 11 breaks, respectively. An additional 5.34 Mb of sequence (without gaps) was localized using the markers located on 13 additional scaffolds that were not previously included in version 1.0. Of these 13 additional scaffolds located on seven chromosomes, three were positioned on Chr2 and Chr3, two on Chr4 and Chr8, and one on Chr1, Chr5 and Chr9. All additional scaffolds were located at a similar genetic position on both parental maps, validating their position (Table S11).

The newly integrated version 2.0 of *E. grandis* resulting from these improvements contained 304 map joins made on 315 scaffolds to form the 11 chromosomes capturing 612.6 Mb (88.6%) of the assembled sequence. All 43 NCRs and the 13 NSRs common to the two parental maps, amounting to 90 scaffolds, were corrected in version 2.0 (see Fig. 4b). Contaminant screening identified scaffolds as unanchored rDNA (14 scaffolds; 374.7 kb), mitochondrion (six scaffolds; 495.5 kb), unanchored repeats (215 scaffolds; 960.7 kb), chloroplast (139 scaffolds; 978.4 kb), and prokaryote (51 scaffolds; 317.7 kb). Version 2.0 of the *E. grandis* release consists of the 11 integrated chromosomes and remaining 4932 scaffolds for a total of 4943 scaffolds. The total size of version 2.0 is 691.3 Mb (640.4 Mb of sequence and 7.4% scaffold gaps). Contig L50 was 67.2 kb (32 835 total contigs) and scaffold L50 was 57.5 Mb. Verification of the completeness of the genome was performed by aligning 1 634 940 EST sequences to version 2.0 using BLAT with default parameters. A total of 97.2% of the ESTs aligned at > 90% identity and 85% coverage, with only 0.93% not being found.

Discussion

Performance of the SNP array

This study is the first to report on the designing of an Infinium SNP array for *Eucalyptus*. Our approach for SNP detection was based on whole-genome resequencing of the two parental genotypes of a mapping population with high sequencing depth

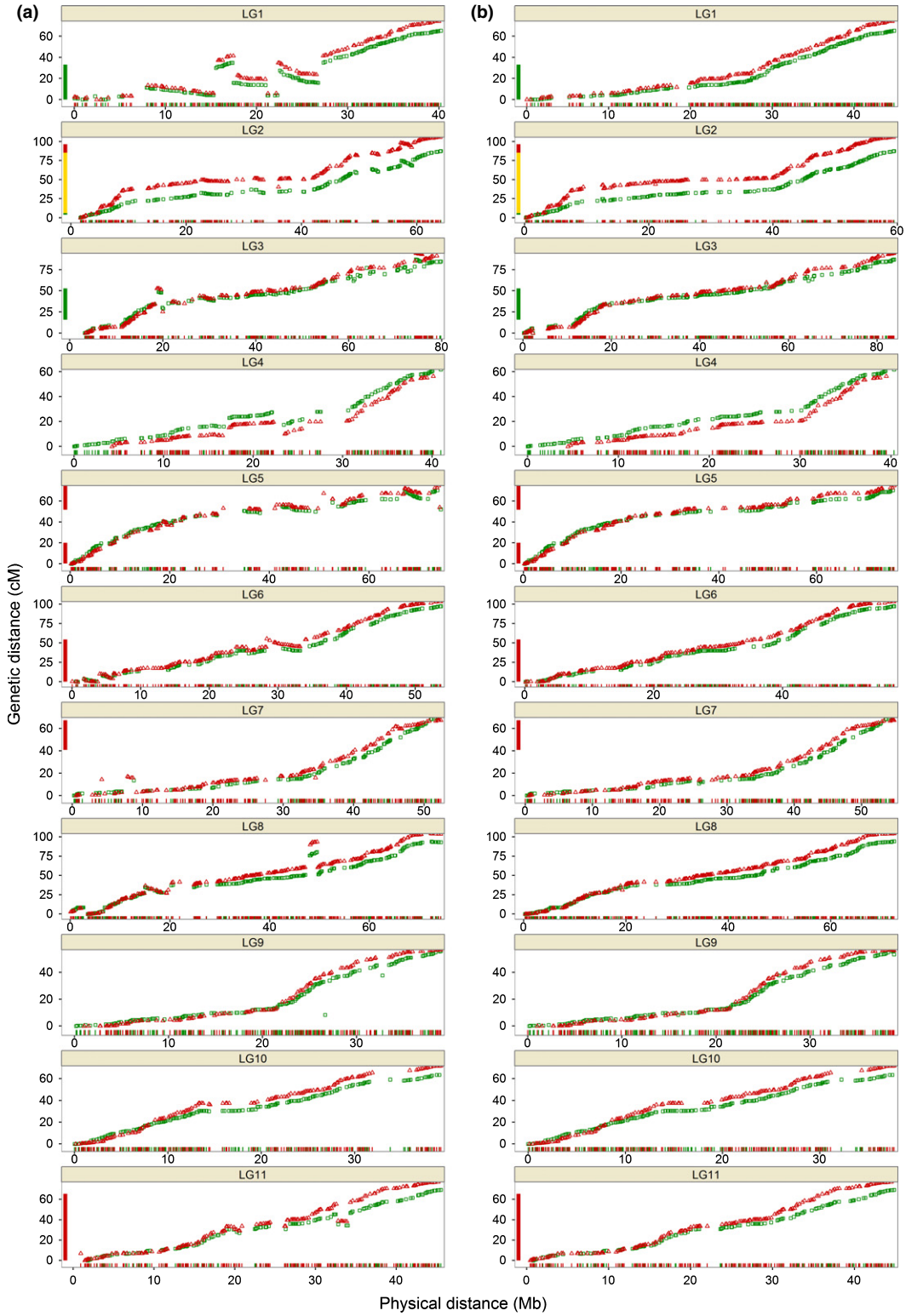

Fig. 4 Physical position (in Mb) and the genetic location (in cM) for the single nucleotide polymorphisms (SNPs) mapped on *Eucalyptus grandis* (green squares) and *Eucalyptus urophylla* (red triangles) framework maps. The physical position on version 1.0 of the BRASUZ1 genome sequence is presented in (a) and in (b) for version 2.0. Segregation distortion regions ($P_{adjust} < 0.01$) are represented by vertical lines along the *y*-axis, in green for *E. grandis*, in red for *E. urophylla* and in yellow when the two parents are involved. Only markers with a conserved chromosomal assignment between genetic and physical maps are represented.

(≈40X). Stringent selection criteria resulted in a high genotyping SR, that is, 84.4% of the genotyped SNPs were polymorphic in the studied pedigree. Such a high SR was also reported in *P. trichocarpa* using a similar resequencing approach, with an SR of 93.1% in natural populations (Geraldes *et al.*, 2013). Chagné *et al.* (2012) reported an SR of 72.2% in *Malus pumila*

Table 3 Features of the noncollinear regions (NCRs) and nonsyntenic regions (NSRs) between genetic framework maps for *Eucalyptus grandis* (*E.g*) and *E. urophylla* (*E.u*), and the BRASUZ1 genome sequence

			E.g	*E.u*
NCRs	All	Number of regions	49	51
		Number of SNPs	393	382
		% of SNPs	21.5	21.5
		% of genome size	13.9	15.3
	Common	Number of regions	43	43
		% of regions	87.8	84.3
		Number of SNPs	378	358
		% of SNPs	20.6	20.2
		% of genome size	13.5	14.4
NSRs	All	Number of regions	20	20
		Number of SNPs	45	41
		% of SNPs	2.5	2.3
		% of genome size	1	0.7
		LG final	1, 3, 5, 6, 7, 8, 11	1, 2, 3, 5, 6, 7, 8, 9
	Common	Number of regions	13	13
		% of regions	65	65
		Number of SNPs	35	33
		% of SNPs	1.9	1.9
		% of genome size	0.8	0.7
		LG final	1, 3, 5, 6, 7, 8	

SNP, single nucleotide polymorphism; LG, linkage group.
Common NCRs or NSRs are supported by both parental maps.

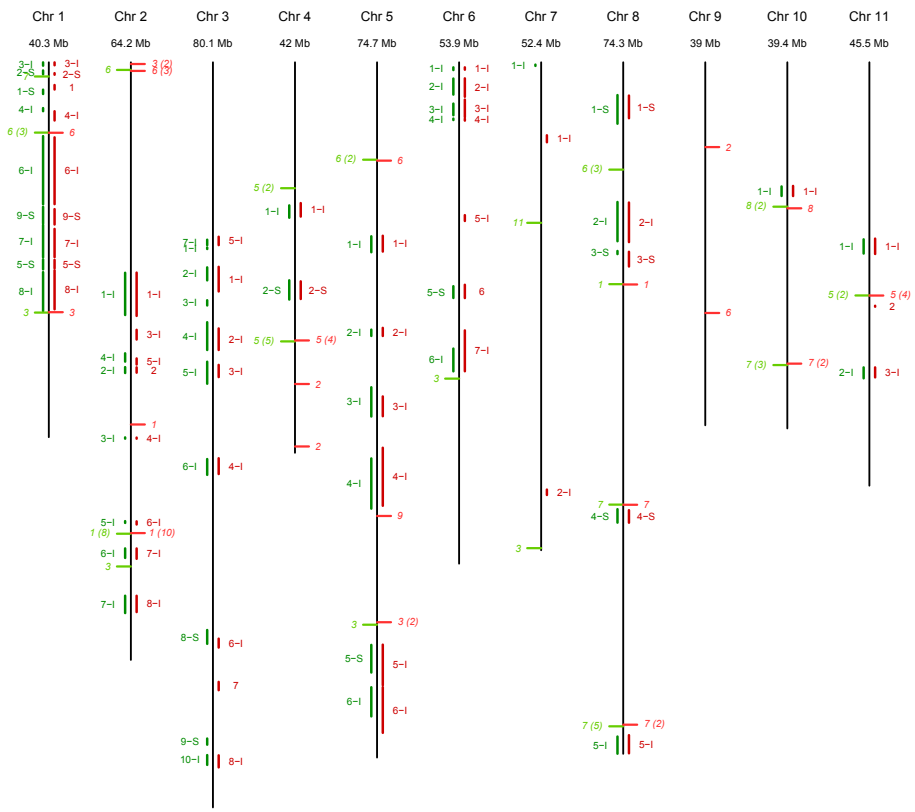

Fig. 5 Physical location of noncollinear regions (NCRs) and nonsyntenic regions (NSRs) on the *BRASUZ1* genome sequence (v1.0). Vertical black lines represent the physical skeleton of the 11 chromosomes (Chr). Discordant regions between physical and genetic maps are in green for *Eucalyptus grandis* (left) and red for *Eucalyptus urophylla* (right). Vertical lines on each scaffold represent NCRs. They are numbered in the same way as in Supporting Information Table S9. I, inverted NCRs; S, displaced NCRs. Horizontal tick bars indicate NSRs, followed by the linkage group number onto which the NSR is mapped (*in italics*). The number in brackets indicates the number of markers involved in the NSR (if > 1).

despite an additional validation step on a subset of SNPs with the GoldenGate technology. This contrast with the lower values generally obtained in nonmodel forest tree species using the Infinium technology and mainly SNPs detected from RNA-seq data. For instance, Pavy *et al.* (2013) designed two SNP arrays in *Picea glauca*, and reported SRs of 55.8% and 67.6%. Similarly, Howe *et al.* (2013) obtained an SR of 72.5% in *Pseudotsuga menziesii*. This disparity between studies might be

explained by various factors, such as the stringency of thresholds at different stages of SNP detection and selection, the nature of the primary data (RNA versus whole-genome DNA) and the genetic proximity between the discovery panel and the genotyped populations, as illustrated by Chancerel *et al.* (2013) in *Pinus pinaster*.

Linkage mapping accuracy

The two framework maps were built independently with SNPs segregating 1 : 1. In contrast to other high-density mapping studies in *Eucalyptus* (Hudson *et al.*, 2012a; Kullan *et al.*, 2012; Petroli *et al.*, 2012), we did not combine parental maps in a consensus map because of the loss of parental-specific features and the loss of accuracy introduced by combining markers with different segregation ratios, resulting in a loss of information (Ritter *et al.*, 1990). The use of markers with different segregation types (mandatory for combining parental maps) indeed reduces mapping accuracy, as illustrated by the inversion of 9.3% of inter-cross markers between the two parental maps obtained using both test-cross and inter-cross markers. Additional factors could also lead to poorer order and thereby bias the resolution of a genetic map, for example inconsistencies between individual maps such as local reordering and/or large displacements (Jackson *et al.*, 2008; Wu *et al.*, 2011; Ronin *et al.*, 2012).

The two parental framework maps provided a comprehensive view of the *Eucalyptus* genome, with 98.2% and 97.2% coverage of the 11 chromosomes and an average marker interval of 321 and 332 kb for *E. grandis* and *E. urophylla*, respectively. The corresponding map lengths were 821.7 and 885.9 cM for *E. grandis* and *E. urophylla*, respectively. Previous genetic size estimates of the genome in different *Eucalyptus* species ranged from 632 to 1815 cM (Table S1), with the lower and higher limits obtained with low-density and/or nonsaturated genetic maps (Brondani *et al.*, 2006; Rocha *et al.*, 2007). Compared with recent studies (Hudson *et al.*, 2012a,b; Kullan *et al.*, 2012; Petroli *et al.*, 2012) which used a DArT array to build dense genetic maps, our estimates of the *Eucalyptus* genome size showed a reduction of at least 11% (*E. grandis*) and 20% (*E. urophylla*). Moreover, this reduction must have been slightly underestimated as we used the Haldane mapping function which is known to give longer maps compared with the Kosambi mapping function (Kosambi, 1943). Indeed, previously published dense genetic maps were built using the regression mapping algorithm of JoinMap in combination with the Kosambi mapping function. Although the mapping algorithm and the mapping function play a role in estimating the genetic map length, the number of recombination events per meiosis captured in the mapping population is one of the main drivers of mapping accuracy (Ferreira *et al.*, 2006). In our study, the number of recombination events was higher than for the most recent ultra-dense genetic maps in *Eucalyptus* (Table S1). Thus, the two parental framework maps presented in this study provided one of the most accurate estimates of the genetic map length for *E. grandis* and *E. urophylla*, which is of major importance for physical versus genetic size analysis.

To confirm the robustness of the two parental genetic maps generated here, we compared the results obtained with Record to those obtained with JoinMap, one of the most widely used software for genetic map construction (Van Ooijen, 2011a). The total map lengths were similar, being 821.1 and 821.7 cM for *E. grandis* and 884.9 and 885.9 cM for *E. urophylla* for JoinMap and Record, respectively. The SNP order was highly similar, with 99.2% and 99.6% of SNPs exhibiting the same order between the two algorithms for *E. grandis* and *E. urophylla*, respectively (Table S5). Inversions only occurred with tightly linked SNPs with a maximum distance between inverted SNPs of 0.1 cM, confirming the robustness of the framework maps. Moreover, these results were corroborated by highly reliable orders obtained for all the 12 000 maps based on the resampling of F1 offspring, with no more than 1.2% of inverted bins even with a small sample size (100).

Segregation distortion regions

Linkage mapping analysis was carried out with distorted SNPs, not only to maintain high genome coverage in such regions of the genome but also because segregation distortion has little or no effect on mapping accuracy (Hackett & Broadfoot, 2003; Hudson *et al.*, 2012b), which was confirmed in our study by the fact that the same marker order was found for LGs displaying different levels of segregation distortion between parental maps (e.g. LG1 and 11). Moreover, the same marker order was found between different samples of individuals, with or without segregation distortion in a given LG (results not show). The number of offspring (1020) and the high level of synteny between physical maps and genetic maps gave us confidence in the nature of the distortion, that is, evidence for biological causes rather than a technical bias in the genotyping process.

The relatively large proportion of SDRs highlighted in our study (21% (*E. grandis*) and 30.9% (*E. urophylla*) of the mapped SNPs) was also reported in interspecific crosses (Myburg *et al.*, 2003; Brondani *et al.*, 2006; Kullan *et al.*, 2012) and to a lesser extent in intraspecific crosses (Thamarus *et al.*, 2002; Freeman *et al.*, 2006; Hudson *et al.*, 2012b). Although the validation of SDRs between studies is limited by a lack of common markers, one major SDR located on LG2 was also found (thanks to microsatellite loci with a known sequence) in another cross between *E. grandis* and *E. urophylla* (Brondani *et al.*, 2006) and in an inter-provenance cross of *E. globulus* (Freeman *et al.*, 2006), suggesting the presence of causal genes involved in hybrid incompatibility. SDRs were found be related to different physiological and genetic factors, such as pollen-tube competition (Arnold *et al.*, 1993; Rahmé *et al.*, 2009; Zhang *et al.*, 2011); pollen–pistil incompatibility, which was found to increase with phylogenetic distance in *Eucalyptus* species (Gore *et al.*, 1990; Ellis *et al.*, 1991) and in the closely related *Corymbia* species (Dickinson *et al.*, 2012); negative epistatic interactions among alleles (Törjék *et al.*, 2006; Bikard *et al.*, 2009); and lethal genes in a homozygous state (Maheshwari & Barbash, 2011). Hybridization enhances some of these processes (Rieseberg & Carney, 1998; Potts & Dungey, 2004; Maheshwari & Barbash, 2011), thus increasing segregation

distortion in hybrids, as found in here and in other genetic mapping studies on *Eucalyptus*.

Validation and improvement of the *Eucalyptus* BRASUZ1 genome assembly

The comparison between parental genetic maps and the reference genome highlighted the high conservation of genome structure between the two species (*E. grandis* and *E. urophylla*). No inconsistencies involving a large number of SNPs were detected between the two maps. This collinearity and synteny between *Eucalyptus* species had already been reported using genetic (Myburg *et al.*, 2003; Hudson *et al.*, 2012b) and physical maps (Myburg *et al.*, 2014). This is consistent with the small phylogenic distance between these two interfertile species of the *Symphyomyrtus* subgenus (Potts & Dungey, 2004; Steane *et al.*, 2011). Moreover, their genomes were found to be close in terms of physical size (Grattapaglia & Bradshaw, 1994; Praça *et al.*, 2009) and in terms of genome content, as shown by the robust alignment of *E. urophylla* sequences on the BRASUZ1 genome. This closeness of genome structure was also reported between *E. grandis* and *E. globulus*, two species belonging to different sections: *Latoangulatae* and *Maidenaria*. The specific genomic regions of these two species were found to be distributed along the genome and mainly related to non-transposable element changes (Myburg *et al.*, 2014).

In our study, the use of two independent and highly collinear genetic maps enabled cross-validation in the comparison between genetic maps and the reference genome sequence. Our results confirmed the robust assembly of most of the BRASUZ1 genome. Indeed, collinear regions accounted for *c.* 85% of the genome size. This first version was assembled using a DArT consensus genetic map (Kullan *et al.*, 2012). Initially, only 78% of the genome sequence was organized into 11 chromosomes because of a bias in the distribution of DArT markers along the genome. To improve the completeness of the genome, additional steps were performed including the use of other genetic maps to independently validate scaffold order and orientation (Myburg *et al.*, 2014). Although inconsistencies between physical and genetic orders were reported for different LGs (e.g. 1 and 4) in a previous study (Petroli *et al.*, 2012), they were attributed to lower marker quality alignments rather than wrong scaffold assembly. The NCRs and NSRs highlighted by the high-resolution genetic maps in this study confirmed the wrong scaffold assembly hypothesis. Therefore, all NCRs and NSRs corroborated by both parental maps (accounting for nearly 14% and 1% of the genome, respectively) were corrected in version 2.0 of the BRASUZ1 genome. Moreover, the completeness of the genome was increased by the inclusion of 13 unanchored scaffolds amounting to 5.34 Mb (excluding gaps). As reported previously (Petroli *et al.*, 2012; Myburg *et al.*, 2014), the majority of small unanchored scaffolds (< 20 kb) probably correspond to already assembled parts of the genome (i.e. alternative haplotypes of the sequenced genotype attributable to the genomic region of high heterozygosity). Thus, the 11 chromosomes of the *E. grandis* genome v2.0 accounted for nearly 95% of the estimated genome size.

Conclusions and prospects

Our SNP array maximizing genome coverage with 6000 SNPs evenly spaced along the 11 chromosomes of the BRASUZ1 genome resulted in two high-resolution genetic linkage maps, providing a framework for future map-based cloning activities of major-effect QTLs, and an improved version (2.0) of the *E. grandis* genome assembly available on the Phytozome 10 (http://genome.jgi.doe.gov/pages/dynamicOrganismDownload.jsf?organism=Egrandis). These two results provide a robust basis for further studies on the recombination rate (cM/Mb) and its relationship with genome features in *Eucalyptus*.

Acknowledgements

The authors would like to thank CRDPI in the Republic of the Congo. The authors also thank Matthieu Falque who provided the R script for Fig. 5. This study was supported by grants from FEDER (ABIOGEN project, no. Presage 32973) and ERANET Plant KBBE (ANR (FR)-10-KBBE-0007, '34Joule'). J.B. received a PhD fellowship from the BIOS department of CIRAD.

References

Arnold ML, Hamrick JL, Bennett BD. 1993. Interspecific pollen competition and reproductive isolation in Iris. *Journal of Heredity* **84**: 13–16.

Benjamini Y, Yekutieli D. 2001. The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics* **29**: 1165–1188.

Bikard D, Patel D, Le Metté C, Giorgi V, Camilleri C, Bennett MJ, Loudet O. 2009. Divergent evolution of duplicate genes leads to genetic incompatibilities within *A. thaliana*. *Science* **323**: 623–626.

Botstein D, White RL, Skolnick M, Davis RW. 1980. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American Journal of Human Genetics* **32**: 314.

Brondani R, Brondani C, Tarchini R, Grattapaglia D. 1998. Development, characterisation and mapping of microsatellite markers in *Eucalyptus grandis* and *E. urophylla*. *Theoretical and Applied Genetics* **97**: 816–827.

Brondani R, Williams E, Brondani C, Grattapaglia D. 2006. A microsatellite-based consensus linkage map for species of *Eucalyptus* and a novel set of 230 microsatellite markers for the genus. *BMC Plant Biology* **6**: 20.

Burt DW. 2002. Comparative mapping in farm animals. *Briefings in Functional Genomics & Proteomics* **1**: 159–168.

Chagné D, Crowhurst RN, Troggio M, Davey MW, Gilmore B, Lawley C, Vanderzande S, Hellens RP, Kumar S, Cestaro A *et al.* 2012. Genome-wide SNP detection, validation, and development of an 8K SNP array for apple. *PLoS ONE* **7**: e31745.

Chancerel E, Lamy J-B, Lesur I, Noirot C, Klopp C, Ehrenmann F, Boury C, Provost GL, Label P, Lalanne C *et al.* 2013. High-density linkage mapping in a pine tree reveals a genomic region associated with inbreeding depression and provides clues to the extent and distribution of meiotic recombination. *BMC Biology* **11**: 50.

Cheema J, Dicks J. 2009. Computational approaches and software tools for genetic linkage map estimation in plants. *Briefings in Bioinformatics* **10**: 595–608.

Choi H-K, Mun J-H, Kim D-J, Zhu H, Baek J-M, Mudge J, Roe B, Ellis N, Doyle J, Kiss GB *et al.* 2004. Estimating genome conservation between crop and model legume species. *Proceedings of the National Academy of Sciences, USA* **101**: 15289–15294.

Collard B, Mace E, McPhail M, Wenzl P, Cakir M, Fox G, Poulsen D, Jordan D. 2009. How accurate are the marker orders in crop linkage maps generated from large marker datasets? *Crop and Pasture Science* **60**: 362–372.

Crismani W, Girard C, Mercier R. 2013. Tinkering with meiosis. *Journal of Experimental Botany* **64**: 55–65.

Dickinson GR, Lee DJ, Wallace HM. 2012. The influence of pre- and post-zygotic barriers on interspecific *Corymbia* hybridization. *Annals of Botany* **109**: 1215–1226.

Doyle J, Doyle J. 1990. Isolation of plant DNA from fresh tissue. *Focus* **12**: 13–15.

Ellis MF, Sedgley M, Gardner JA. 1991. Interspecific pollen–pistil interaction in *Eucalyptus* L'Hér. (Myrtaceae): the effect of taxonomic distance. *Annals of Botany* **68**: 185–194.

Endelman JB, Plomion C. 2014. LPmerge: an R package for merging genetic maps by linear programming. *Bioinformatics* **30**: 1623–1624.

Ferreira A, da Silva MF, Silva L, Cruz CD. 2006. Estimating the effects of population size and type on the accuracy of genetic maps. *Genetics and Molecular Biology* **29**: 187–192.

Freeman JS, Potts BM, Shepherd M, Vaillancourt RE. 2006. Parental and consensus linkage maps of *Eucalyptus globulus* using AFLP and microsatellite markers. *Silvae Genetica* **55**: 202–217.

Ganal MW, Durstewitz G, Polley A, Bérard A, Buckler ES, Charcosset A, Clarke JD, Graner E-M, Hansen M, Joets J *et al.* 2011. A large maize (*Zea mays* L.) SNP genotyping array: development and germplasm genotyping, and genetic mapping to compare with the B73 reference genome. *PLoS ONE* **6**: e28334.

Geraldes A, DiFazio SP, Slavov GT, Ranjan P, Muchero W, Hannemann J, Gunter LE, Wymore AM, Grassa CJ, Farzaneh N *et al.* 2013. A 34K SNP genotyping array for *Populus trichocarpa*: design, application to the study of natural populations and transferability to other *Populus* species. *Molecular Ecology Resources* **13**: 306–323.

Gion J-M, Rech P, Grima-Pettenati J, Verhaegen D, Plomion C. 2000. Mapping candidate genes in *Eucalyptus* with emphasis on lignification genes. *Molecular Breeding* **6**: 441–449.

Gore PL, Potts BM, Volker PW, Megalos J. 1990. Unilateral cross-incompatibility in *Eucalyptus*: the case of hybridisation between *E. globulus* and *E. nitens*. *Australian Journal of Botany* **38**: 383–394.

Grattapaglia D, Bradshaw HD Jr. 1994. Nuclear DNA content of commercially important *Eucalyptus* species and hybrids. *Canadian Journal of Forest Research* **24**: 1074–1078.

Grattapaglia D, Sederoff R. 1994. Genetic linkage maps of *Eucalyptus grandis* and *Eucalyptus urophylla* using a pseudo-testcross mapping strategy and RAPD markers. *Genetics* **137**: 1121–1137.

Grattapaglia D, Silva-Junior O, Kirst M, de Lima B, Faria D, Pappas G. 2011. High-throughput SNP genotyping in the highly heterozygous genome of *Eucalyptus*: assay success, polymorphism and transferability across species. *BMC Plant Biology* **11**: 65.

Hackett CA, Broadfoot LB. 2003. Effects of genotyping errors, missing values and segregation distortion in molecular marker data on the construction of linkage maps. *Heredity* **90**: 33–38.

Haldane JBS. 1919. The combination of linkage values, and the calculation of distance between the loci of linked factors. *Journal of Genetics* **8**: 299–309.

Howe GT, Yu JB, Knaus B, Cronn R, Kolpak S, Dolan P, Lorenz WW, Dean JFD. 2013. A SNP resource for Douglas-fir: *de novo* transcriptome assembly and SNP detection and validation. *BMC Genomics* **14**: UNSP 137.

Hudson CJ, Freeman JS, Kullan ARK, Petroli CD, Sansaloni CP, Kilian A, Detering F, Grattapaglia D, Potts BM, Myburg AA *et al.* 2012a. A reference linkage map for eucalyptus. *BMC Genomics* **13**: 240.

Hudson CJ, Kullan ARK, Freeman JS, Faria DA, Grattapaglia D, Kilian A, Myburg AA, Potts BM, Vaillancourt RE. 2012b. High synteny and colinearity among *Eucalyptus* genomes revealed by high-density comparative genetic mapping. *Tree Genetics & Genomes* **8**: 339–352.

Jackson BN, Schnable PS, Aluru S. 2008. Consensus genetic maps as median orders from inconsistent sources. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **5**: 161–171.

Jaffe DB, Butler J, Gnerre S, Mauceli E, Lindblad-Toh K, Mesirov JP, Zody MC, Lander ES. 2003. Whole-genome sequence assembly for mammalian genomes: Arachne 2. *Genome Research* **13**: 91–96.

Kaul S, Koo HL, Jenkins J, Rizzo M, Rooney T, Tallon LJ, Feldblyum T, Nierman W, Benito MI, Lin XY *et al.* 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**: 796–815.

Kent WJ. 2002. BLAT – the BLAST-like alignment tool. *Genome Research* **12**: 656–664.

Kole C. 2007. *Genome mapping and molecular breeding in plants: forest trees*. Berlin, Germany: Springer.

Koornneef M, Vaneden J, Hanhart CJ, Stam P, Braaksma FJ, Feenstra WJ. 1983. Linkage map of *Arabidopsis thaliana*. *Journal of Heredity* 74: 265–272.

Kosambi DD. 1943. The estimation of map distances from recombination values. *Annals of Human Genetics* **12**: 172–175.

Krutovsky KV, Troggio M, Brown GR, Jermstad KD, Neale DB. 2004. Comparative mapping in the Pinaceae. *Genetics* **168**: 447–461.

Kullan ARK, van Dyk M, Jones N, Kanzler A, Bayley A, Myburg A. 2012. High-density genetic linkage maps with over 2,400 sequence-anchored DArT markers for genetic dissection in an F2 pseudo-backcross of *Eucalyptus grandis* × *E. urophylla*. *Tree Genetics & Genomes* **8**: 163–175.

Kumar S, Banks TW, Cloutier S. 2012. SNP discovery through next-generation sequencing and its applications. *International Journal of Plant Genomics* **2012**: 15.

Lander ES, Int Human Genome Sequencing C, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M *et al.* 2001. Initial sequencing and analysis of the human genome. *Nature* **409**: 860–921.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**: 1754–1760.

Mackay TFC, Stone EA, Ayroles JF. 2009. The genetics of quantitative traits: challenges and prospects. *Nature Reviews Genetics* **10**: 565–577.

Maheshwari S, Barbash DA. 2011. The genetics of hybrid incompatibilities. *Annual Review of Genetics* **45**: 331–355.

Margarido GRA, Souza AP, Garcia AAF. 2007. OneMap: software for genetic mapping in outcrossing species. *Hereditas* **144**: 78–79.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M *et al.* 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* **20**: 1297–1303.

Mollinari M, Margarido GRA, Vencovsky R, Garcia AAF. 2009. Evaluation of algorithms used to order markers on genetic maps. *Heredity* **103**: 494–502.

Myburg AA, Grattapaglia D, Tuskan GA, Hellsten U, Hayes RD, Grimwood J, Jenkins J, Lindquist E, Tice H, Bauer D *et al.* 2014. The genome of *Eucalyptus grandis*. *Nature* **510**: 356–362.

Myburg AA, Griffin AR, Sederoff RR, Whetten RW. 2003. Comparative genetic linkage maps of *Eucalyptus grandis*, *Eucalyptus globulus* and their F1 hybrid based on a double pseudo-backcross mapping approach. *Theoretical and Applied Genetics* **107**: 1028–1042.

Neale DB, Kremer A. 2011. Forest tree genomics: growing resources and applications. *Nature Reviews Genetics* **12**: 111–122.

Nelson JC. 2005. Methods and software for genetic mapping. In: Meksem K, Kahl G, eds. *The handbook of plant genome mapping*. Weinheim, Germany: Wiley-VCH Verlag GmbH & Co., 53–74.

Neves L, Mc Mamani E, Alfenas A, Kirst M, Grattapaglia D. 2011. A high-density transcript linkage map with 1,845 expressed genes positioned by microarray-based single feature polymorphisms (SFP) in *Eucalyptus*. *BMC Genomics* **12**: 189.

Novaes E, Drost DR, Farmerie WG, Pappas GJ, Grattapaglia D, Sederoff RR, Kirst M. 2008. High-throughput gene and SNP discovery in *Eucalyptus grandis*, an uncharacterized genome. *BMC Genomics* **9**: 312.

Os H, Stam P, Visser RF, Eck H. 2005. RECORD: a novel method for ordering loci on a genetic linkage map. *Theoretical and Applied Genetics* **112**: 30–40.

Pavy N, Gagnon F, Rigault P, Blais S, Deschenes A, Boyle B, Pelgas B, Deslauriers M, Clement S, Lavigne P *et al.* 2013. Development of high-density SNP genotyping arrays for white spruce (*Picea glauca*) and transferability to subtropical and nordic congeners. *Molecular Ecology Resources* **13**: 324–336.

Peirce JL, Broman KW, Lu L, Williams RW. 2007. A simple method for combining genetic mapping data from multiple crosses and experimental designs. *PLoS ONE* **2**: e1036.

Petroli CD, Sansaloni CP, Carling J, Steane DA, Vaillancourt RE, Myburg AA, da Silva OB Jr, Pappas GJ Jr, Kilian A, Grattapaglia D. 2012. Genomic

characterization of DArT markers based on high-density linkage analysis and physical mapping to the *Eucalyptus* genome. *PLoS ONE* **7**: e44684.
Potts BM, Dungey HS. 2004. Interspecific hybridization of *Eucalyptus*: key issues for breeders and geneticists. *New Forests* **27**: 115–138.
Praça MM, Carvalho CR, Novaes CRDB. 2009. Nuclear DNA content of three *Eucalyptus* species estimated by flow and image cytometry. *Australian Journal of Botany* **57**: 524–531.
Price AH. 2006. Believe it or not, QTLs are accurate! *TRENDS in Plant Science* **11**: 213–216.
Rahmé J, Widmer A, Karrenberg S. 2009. Pollen competition as an asymmetric reproductive barrier between two closely related *Silene* species. *Journal of Evolutionary Biology* **22**: 1937–1943.
Rieseberg LH, Carney SE. 1998. Plant hybridization. *New Phytologist* **140**: 599–624.
Ritter E, Gebhardt C, Salamini F. 1990. Estimation of recombination frequencies and construction of RFLP linkage maps in plants from crosses between heterozygous parents. *Genetics* **125**: 645–654.
Rocha RB, Barros EG, Cruz CD, Rosado AM, de Araujo EF. 2007. Mapping of QTLs related with wood quality and developmental characteristics in hybrids (*Eucalyptus grandis* * *Eucalyptus urophylla*). *Revista Arvore* **31**: 13–24.
Ronin Y, Mester D, Minkov D, Belotserkovski R, Jackson BN, Schnable PS, Aluru S, Korol A. 2012. Two-phase analysis in consensus genetic mapping. *G3: Genes – Genomes – Genetics* **2**: 537–549.
Ronin Y, Mester D, Minkov D, Korol A. 2010. Building reliable genetic maps: different mapping strategies may result in different maps. *Natural Science* **2**: 576–589.
Salvi S, Tuberosa R. 2007. Cloning QTLs in plants. In: Varshney R, Tuberosa R, eds. *Genomics-assisted crop improvement*. Dordrecht, the Netherlands: Springer, 207–225.
Sansaloni CP, Petroli CD, Carling J, Hudson CJ, Steane DA, Myburg AA, Grattapaglia D, Vaillancourt RE, Kilian A. 2010. A high-density Diversity Arrays Technology (DArT) microarray for genome-wide genotyping in *Eucalyptus*. *Plant Methods* **6**: 16.
Schlotterer C. 2004. The evolution of molecular markers – just a matter of fashion? *Nature Reviews Genetics* **5**: 63–69.
Shen R, Fan J-B, Campbell D, Chang W, Chen J, Doucet D, Yeakley J, Bibikova M, Wickham Garcia E, McBride C *et al.* 2005. High-throughput SNP genotyping on universal bead arrays. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* **573**: 70–82.
Stam P. 1993. Construction of integrated genetic linkage maps by means of a new computer package: Join Map. *Plant Journal* **3**: 739–744.
Steane DA, Nicolle D, McKinnon GE, Vaillancourt RE, Potts BM. 2002. Higher-level relationships among the eucalypts are resolved by ITS-sequence data. *Australian Systematic Botany* **15**: 49–62.
Steane DA, Nicolle D, Sansaloni CP, Petroli CD, Carling J, Kilian A, Myburg AA, Grattapaglia D, Vaillancourt RE. 2011. Population genetic analysis and phylogeny reconstruction in *Eucalyptus* (Myrtaceae) using high-throughput, genome-wide genotyping. *Molecular Phylogenetics and Evolution* **59**: 206–224.
Sturtevant AH. 1913. The linear arrangement of six sex-linked factors in *Drosophila*, as shown by their mode of association. *Journal of Experimental Zoology* **14**: 43–59.
Thamarus KA, Groom K, Murrell J, Byrne M, Moran GF. 2002. A genetic linkage map for *Eucalyptus globulus* with candidate loci for wood, fibre, and floral traits. *Theoretical and Applied Genetics* **104**: 379–387.
Törjék O, Witucka-Wall H, Meyer R, Korff M, Kusterer B, Rautengarten C, Altmann T. 2006. Segregation distortion in Arabidopsis C24/Col-0 and Col-0/C24 recombinant inbred line populations is due to reduced fertility caused by epistatic interaction of two loci. *Theoretical and Applied Genetics* **113**: 1551–1561.
Tuskan GA, DiFazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A *et al.* 2006. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* **313**: 1596–1604.
Van Ooijen JW. 2011a. *JoinMap 4.1, software for the calculation of genetic linkage maps in experimental populations*. Wageningen, the Netherlands: Kyazma BV.
Van Ooijen JW. 2011b. Multipoint maximum likelihood mapping in a full-sib family of an outbreeding species. *Genetics Research* **93**: 343–349.
Wu Y, Close TJ, Lonardi S. 2011. Accurate construction of consensus genetic maps via integer linear programming. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **8**: 381–394.
Wurschum T. 2012. Mapping QTL for agronomic traits in breeding populations. *Theoretical and Applied Genetics* **125**: 201–210.
Yelina NE, Choi K, Chelysheva L, Macaulay M, de Snoo B, Wijnker E, Miller N, Drouaud J, Grelon M, Copenhaver GP *et al.* 2012. Epigenetic remodeling of meiotic crossover frequency in *Arabidopsis thaliana* DNA methyltransferase mutants. *PLoS Genetics* **8**: e1002844.
Zhang X-L, Li X-X, Ni L-Y, Guo Y-H. 2011. When interspecific prezygotic barriers break down: hybridization between two *Potamogeton* species (*P. wrightii* and *P. perfoliatus*) and comparison of the artificial and natural hybrids. *Plant Systematics and Evolution* **295**: 119–128.
Zhang Z, Schwartz S, Wagner L, Miller W. 2000. A greedy algorithm for aligning DNA sequences. *Journal of Computational Biology* **7**: 203–214.

Supporting Information

Additional supporting information may be found in the online version of this article.

Fig. S1 Examples of SNP clusters obtained with GenomeStudio software.

Fig. S2 Distribution of mapped reads on the 11 chromosomes of the BRASUZ1 genome sequence.

Fig. S3 Distribution of selected SNPs along the 11 chromosomes of the BRASUZ1 genome sequence.

Fig. S4 Pattern of SNP segregation along the 11 linkage groups of the *Eucalyptus* genome.

Fig. S5 Comparison of the genetic location of SNPs segregating 1 : 2 : 1 (Set 3 in the Materials and Methods section) between the two parental maps over the 11 linkage groups.

Table S1 Comprehensive list of genetic linkage maps established for *Eucalyptus* species

Table S2 List of successful SNPs

Table S3 Detailed SNP genotyping results per scaffold and SNP set (Set 1: SNPs segregating 1 : 1 in *Eucalyptus grandis*; Set 2: SNPs segregating 1 : 1 in *Eucalyptus urophylla*; Set 3: SNPs segregating 1 : 2 : 1 in both species)

Table S4 Location of monomorphic regions in the *Eucalyptus* genome and the genotyping results for the associated SNPs

Table S5 Framework linkage map features based on test-cross markers only (1 : 1 segregation) for *Eucalyptus grandis* (*E.g*) and *Eucalyptus urophylla* (*E.u*). Two algorithms ML (JoinMap) and Record (Onemap) are compared

Table S6 Linkage map features based on test-cross (1 : 1 segregation) and inter-cross (1 : 2 : 1) markers for *Eucalyptus grandis* (*E.g*) and *Eucalyptus urophylla* (*E.u*)

Table S7 Characteristics of the four marker density (MD) classes based on framework maps for *Eucalyptus grandis* (*E.g*) and *Eucalyptus urophylla* (*E.u*)

Table S8 Averages and standard deviations for 1000 genetic maps established for different modalities of marker density (MD) and sample size (100, 200, and 500)

Table S9 Physical coordinates of noncollinear regions (NCRs) on the BRASUZ1 genome sequence

Table S10 Physical coordinates of the nonsyntenic regions (NSRs) on the BRASUZ1 genome sequence

Table S11 Map location of additional scaffolds